Developing Optimal Search Strategies for Detecting Sound Clinical Prediction Studies in MEDLINE

Sharon S.-L. Wong, MSc, Nancy L. Wilczynski, MSc, R. Brian Haynes, MD, PhD, Ravi Ramkissoonsingh, MA, for the Hedges Team
Health Information Research Unit, McMaster University, Hamilton, Ontario, Canada

Abstract

Background: The gaining interest in the use of clinical prediction guides as an aid for helping clinicians make effective front-line decisions, together with the increasing emphasis on evidence-based practice, underscores the need for accurate identification of sound clinical prediction studies. Despite the growing use of clinical prediction guides, little work has been done on identifying optimal literature search filters for retrieving these types of studies. The current study extends our earlier work, on developing optimal search strategies, to include clinical prediction guides.
Objective: To develop optimal search strategies for detecting methodologically sound clinical prediction studies in MEDLINE in the publishing year 2000.
Design: Comparison of the retrieval performance of methodologic search strategies in MEDLINE with a manual review ("gold standard") of each article for each issue of 161 core health care journals for the year 2000.
Methods: 6 experienced research assistants who had been trained and intensively calibrated reviewed all issues of 161 journals for the publishing year 2000. Each article was classified for format, interest, purpose, and methodologic rigor. Search strategies were developed for all purpose categories, including studies of clinical prediction guides.
Main outcome measures: The sensitivity (recall), specificity, precision, and accuracy of single and combinations of search terms.
Results: 39% of original studies classified as a clinical prediction guide were methodologically sound. Combinations of terms reached peak sensitivities of 95%. Compared with the best single term, a three-term strategy increased sensitivity for sound studies by 17% (absolute increase), but with some loss of specificity when sensitivity was maximized. When search terms were combined to optimize sensitivity and specificity, these values reached or were close to 90%.
Conclusion: Several search strategies can enhance the retrieval of sound clinical prediction studies.

Introduction

Clinicians are often required to make predictions based on clinical history, physical examinations, and laboratory results when making a diagnosis or prognosis, examining causes, and choosing treatment options. The use of clinical prediction guides (CPGs, defined as a guide developed to assist in the prediction of some aspect of a disease or condition) as an aid for helping clinicians make effective front-line decisions has been a subject of increasing interest in the past one to two decades [1, 2, 3]. As the literature on CPGs continues to grow, and as the emphasis on evidence-based practice is also rising, the task of identifying relevant and sound CPGs is becoming more essential. This task, however, is a difficult one for several key reasons.

First, over two million new articles are published every year [4], which makes keeping up-to-date with the literature an enormous challenge. This challenge is further complicated by the fact that common electronic bibliographic databases, such as MEDLINE, are very large and include multiple types of articles, many of which are not clinically relevant or of low methodologic quality [5]. Second, a wide range of terminology has been used to depict CPGs, including the terms test, rule, index, equation, scale, score, profile, prognosis, risk estimate, and model. Those involved in maintaining bibliographic databases and who are not familiar with CPG terminology may need to depend on the author's classification, possibly contributing to the inconsistent indexing of these articles. Finally, although methodologic standards for CPGs have been a topic of some interest [1, 2, 6], researchers continue to adopt different methodologic criteria for prediction guides [3, 7].

Methodologic search filters have been developed for improving the accuracy of searching for clinically relevant and sound studies in various contexts [3, 8, 9]. Despite the gaining popularity of CPGs, only one

study has been done to identify optimal search filters for retrieving clinical prediction studies. Ingui and Rogers [3] reported several well-performing search filters for detecting clinical prediction studies that were developed and validated on 4 to 6 selected journals for the years 1991 through 1998. Using somewhat minimal methodologic criteria, they included studies that developed, validated, or evaluated a CPG.

Using more stringent criteria for sound CPG studies and a gold standard based on a much larger journal set (161 journals) for the year 2000, we evaluated as part of a larger study the retrieval properties of single and combination terms for identifying sound CPG studies in MEDLINE. We confined our manual search to the year 2000, having previously established the robustness of empirical search strategies across publication periods (1991 and 2000) [10]. In the 1990s, we developed search filters on a small subset of journals for 4 types of articles (therapy, diagnosis, prognosis, and causation) [11, 12]. In this paper, we report on the extension of this work and provide the information retrieval properties for CPGs.

Methods

The operating characteristics of methodologic search strategies in MEDLINE (accessed using OVID) were compared with a manual review of all articles in each issue of 161 core health care journals for the year 2000. To evaluate MEDLINE strategies designed to retrieve studies meeting basic methodologic criteria for clinical practice, MeSH terms and textwords related to research design features were run as search strategies. These search strategies were treated as diagnostic tests for sound studies, and the manual review of the literature was treated as the "gold standard." The sensitivity (recall), specificity, precision, and accuracy of MEDLINE searches were determined. For example, for each MEDLINE search strategy, sensitivity (recall) was calculated as the proportion of relevant, high-quality citations retrieved; specificity as the proportion of irrelevant, low-quality citations not retrieved; precision as the proportion of retrieved citations that are relevant and high-quality; and accuracy as the proportion of all citations that are correctly classified.

Six research assistants assessed all articles for studies meeting methodologic criteria in 7 purpose categories. All purpose category definitions and associated methodologic rigor have been previously published [7]. Original studies (of interest to the health care of humans) on therapy, diagnosis, prognosis, and causation that tested CPGs were required to meet the methodologic criteria for CPGs as outlined in Table 1.

The 161 journals reviewed in the year 2000 were selected using an iterative process based on recommendations of clinicians and librarians, Science Citation Index Impact Factors, and their ongoing yield of sound and clinically relevant studies and reviews for the disciplines of internal medicine, general medical practice, mental health, and general nursing practice (a list of reviewed journals is available upon request from authors). Research staff underwent training and intensive calibration; inter-rater reliability (assessed by the kappa statistic) for classifying articles according to methodologic criteria was greater than 80% for all purpose categories [7].

To construct a comprehensive set of search terms, a list of MeSH terms and textwords was initially generated, and input was sought from clinicians and librarians in the United States and Canada through interviews with known searchers, requests at meetings and conferences, and requests to the National Library of Medicine. These experts were asked which terms or phrases they used when searching for studies of causation, prognosis, diagnosis, treatment, economics, CPGs, reviews, costs, and a qualitative nature. Terms could be MeSH terms, including publication types (pt), check tags, and subheadings (sh), or textwords (tw) denoting methodology in titles and abstracts of articles. We compiled a list of 5,345 terms (a list of tested terms is available upon request from authors).

Results

49,028 articles were identified after matching the hand search records with the data downloaded from MEDLINE. Of these 234 articles were classified as original studies of CPGs, of which 91 (39%) were methodologically sound. Strategies to retrieve CPGs were developed using the entire database. We did not attempt to validate these strategies in a separate "validation" portion of the database because there were too few methodologically sound CPG articles to allow the database to be split.

The single terms having the best sensitivity, best specificity, and best optimization of sensitivity and specificity for detecting CPG studies in MEDLINE in 2000 are displayed in Table 2. The most notable trade-off was seen when specificity was maximized, which resulted in a clear, but expected, reduction in sensitivity and increase in precision.

The operating characteristics of top-performing two- or three-term strategies are displayed in Table 3. The use of combined terms increased sensitivity. The three-term strategy, "predict:.mp. OR scor:.tw. OR observ:.mp.", yielded the best sensitivity, 95.60%, and had a specificity of 78.70%. Compared with the best sensitivity single term, "predict:.mp." (78.02% sensitivity, 91.30% specificity), the best three-term strategy yielded an absolute increase in sensitivity of 17.58%, but with an absolute loss in specificity of 12.60%.

The two-term strategy, "validation.tw. OR validate.tw.", yielded the best specificity (better than any of the three-term strategies), 99.28%, but with a striking trade-off in sensitivity, which lowered to 53.85%. Yet as expected, when specificity was maximized, precision also improved (reaching 12.28%). Compared with the best sensitivity three-term strategy, this represents an absolute increase in precision of 11.45%.

When search terms were combined to optimize sensitivity and specificity, these values reached or were close to 90%.

Table 1 – Methodologic Rigor Applied for Clinical Prediction Studies

Purpose Category	Methodologic Rigor
Clinical Prediction Guides	Guide is generated in one or more sets of real patients (training set); Guide is validated in an independent set of real patients (test set).

Table 2 – Single Terms with the Best Sensitivity (keeping Specificity ³50%), Best Specificity (keeping Sensitivity ³50%), and Best Optimization of Sensitivity and Specificity (based on absolute [sensitivity-specificity]<15%) for Detecting Clinical Prediction Studies in MEDLINE in 2000

OVID Search Terms	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)
Best Sensitivity* predict:.mp.	78.02	91.30	1.64	91.27
Best Specificity validat:.tw.	57.14	98.69	7.49	98.61
Best Optimization of Sensitivity and Specificity* predict:.mp.	78.02	91.30	1.64	91.27

*The same single term gives the best sensitivity and the best optimization of sensitivity and specificity.

Table 3 –Two- or Three-Term Strategies with the Best Sensitivity (keeping Specificity ³50%), Best Specificity (keeping Sensitivity ³50%), and Best Optimization of Sensitivity and Specificity (based on absolute [sensitivity-specificity]<1%) for Detecting Clinical Prediction Studies in MEDLINE in 2000

OVID Search Strategies	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)
Best Sensitivity predict:.mp. OR scor:.tw. OR observ:.mp.	95.60	78.70	0.83	78.7
Best Specificity validation.tw. OR validate.tw.	53.85	99.28	12.28	99.20
Best Optimization of Sensitivity and Specificity predict:.tw. OR validat:.tw. OR develop.tw.	90.11	89.76	1.61	89.76

Table 4 – Strategies for Detecting Clinical Prediction Guides in MEDLINE from Previous Research [3] Tested in our 2000 Database

OVID Search Strategies - Previously Derived Strategy **Modified Strategy Tested in our Database**	**Sensitivity (%) Original** **2000 Database**	**Specificity (%) Original** **2000 Database**	**Precision (%) Original** **2000 Database**	**Accuracy (%) Original** **2000 Database**
High Sensitivity (Single Term) predict$ predict:.mp.	78.6 78.02	Not reported 91.30	Not reported 1.66	Not reported 91.27
High Specificity (Single Term) Decision Support Techniques/ decision support techniques.sh.	Not reported 3.30	99.5 99.82	Not reported 3.30	Not reported 99.64
High Sensitivity (2-Term Strategy) Predict$ OR Risk$ predict:.mp. OR risk:.mp.	87.5 83.52	78.1 77.37	2.0 0.68	Not reported 77.38
High Specificity (2-Term Strategy) Decision Support Techniques/ AND Predictive Value of Tests/ decision support techniques.sh. AND predictive value of tests.sh.	1.8 2.20	99.96 99.99	20.0 33.33	Not reported 99.81
High Sensitivity and High Specificity (4-Term Strategy) Predict$ OR Validat$ OR Rule$ OR Predictive Value of Tests/ predict:.mp. OR validat:.mp. OR rule:.mp. OR predictive value of tests.sh.	91.1 86.81	93.0 90.01	6.3 1.59	Not reported 90.01

Discussion

We have developed search filters that can enhance the retrieval of clinically relevant and sound CPGs. Clinicians and researchers should examine our filters and determine the most appropriate trade-off between sensitivity and specificity for their searching needs. Those willing to spend the time to sort out irrelevant articles to avoid missing key articles would benefit most from using a highly sensitive strategy. Those wishing to efficiently find several key articles and not requiring an exhaustive collection of relevant articles would benefit most from using a highly specific strategy.

In our strategies, precision was generally low. Maximizing specificity somewhat improved precision, which is expected given that specificity is a major determinant of precision. Low precision in our study was inevitable because MEDLINE is such a large multi-purpose database, and few studies in it are about CPGs. Precision would likely have been greater had we combined our strategies with content terms, or had we tested "and" and "and not" combinations.

We tested several strategies derived from a previous study [3] in our 2000 database (Table 4). When comparing the original performance of previously derived strategies with their performance in our 2000 database, it appeared that sensitivity and specificity was frequently similar, especially for single term strategies. The use of combined terms increased sensitivity. The statistical significance of the absolute differences between the characteristics of previously derived strategies in their original database versus in

our 2000 database could not be determined because numerators and denominators were not provided in the previous study. For the four-term strategy, "predict:.mp. OR validat:.mp. OR rule:.mp. OR predictive value of tests.sh.", precision was 6.3% (95% CI 4.8% to 8.3%) in the original study and 1.59% (CI 1.24% to 1.94%) in our 2000 database (absolute difference 4.71%). When interpreting this absolute difference in precision, it is important to consider that the previous study was based on 4 to 6 journals, and our 2000 database was based on 161 journals (which would naturally imply a much larger number of irrelevant retrievals), further to the differences in publication years accessed.

Further work is needed to develop and validate more sophisticated search strategies (e.g., using "and" and "and not" combinations, or possibly natural language processing) and to determine how well our strategies perform when combined with content and age terms. Although further development of more sophisticated strategies is important for increasing the accuracy of identifying clinically sound CPG evidence, efforts are also needed among researchers and clinicians to establish more uniform agreement on methodologic standards and terminology, which would hopefully lead to more consistent indexing of these studies.

Conclusion

Several search strategies can enhance the retrieval of sound clinical prediction studies, and the optimal trade-off between sensitivity and specificity should be determined according to the searcher's needs.

References

[1] Laupacis A, Sekar N, Steill IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* 1997;277:488-94.

[2] Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. *N Engl J Med* 1985;313:793-9.

[3] Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc* 2001;8:391-7.

[4] Lee KP, Schotland M, Bacchetti P, Bero LA. Association of journal quality indicators with methodological quality of clinical research articles. *JAMA* 2002;287:2805-8.

[5] Haynes RB, Sackett DL, Tugwell P. Problems in handling of clinical and research evidence by medical practitioners. *Arch Intern Med* 1983;143:1971-5.

[6] McGinn T, Randolph A, Richardson S, Sackett D. Clinical prediction guides. *ACP J Club* 1998 Jan-Feb;128:A14-5.

[7] Wilczynski NL, McKibbon KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. *Medinfo* 2001;10(Pt 1):390-3.

[8] Nwosu CR, Khan KS, Chien PF. A two-term MEDLINE search strategy for identifying randomized trials in obstetrics and gynecology. *Obstet Gynecol.* 1998;91:618-22.

[9] Robinson KA, Dickersin K. Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. *Int J Epidemiol.* 2002;31:150-3.

[10] Wilczynski NL, Haynes RB. Robustness of empirical search strategies for clinical content in MEDLINE. *Proc AMIA Symp* 2002;:904-8.

[11] Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Assessment of methodologic search filters in MEDLINE. *Proc Annu Symp Comput Appl Med Care* 1993;:601-5.

[12] Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *J Am Med Inform Assoc* 1994;1:447-58.

Acknowledgments

This research was funded by the National Library of Medicine, USA. The Hedges Team includes Angela Eady, Brian Haynes, Susan Marks, Ann McKibbon, Doug Morgan, Cindy Walker-Dilks, Stephen Walter, Nancy Wilczynski, and Sharon Wong.